Exponential ergodicity of mirror-Langevin diffusions

Sinho Chewi
MIT
schewi@mit.edu

Thibaut Le Gouic
MIT
tlegouic@mit.edu

Chen Lu
MIT
chenl819@mit.edu

Tyler Maunu
MIT
maunut@mit.edu

Philippe Rigollet
MIT
rigollet@mit.edu

Austin Stromme
MIT
astromme@mit.edu

Abstract

Motivated by the problem of sampling from ill-conditioned log-concave distributions, we give a clean non-asymptotic convergence analysis of mirror-Langevin diffusions as introduced in [Zha+20]. As a special case of this framework, we propose a class of diffusions called Newton-Langevin diffusions and prove that they converge to stationarity exponentially fast with a rate which not only is dimension-free, but also has no dependence on the target distribution. We give an application of this result to the problem of sampling from the uniform distribution on a convex body using a strategy inspired by interior-point methods. Our general approach follows the recent trend of linking sampling and optimization and highlights the role of the chi-squared divergence. In particular, it yields new results on the convergence of the vanilla Langevin diffusion in Wasserstein distance.

1 Introduction

Sampling from a target distribution is a central task in statistics and machine learning with applications ranging from Bayesian inference [RC04; DM+19] to deep generative models [Goo+14]. Owing to a firm mathematical grounding in the theory of Markov processes [MT09], as well as its great versatility, Markov Chain Monte Carlo (MCMC) has emerged as a fundamental sampling paradigm. While traditional theoretical analyses are anchored in the asymptotic framework of ergodic theory, this work focuses on finite-time results that better witness the practical performance of MCMC for high-dimensional problems arising in machine learning.

This perspective parallels an earlier phenomenon in the much better understood field of optimization where convexity has played a preponderant role for both theoretical and methodological advances [Nes04; Bub15]. In fact, sampling and optimization share deep conceptual connections that have contributed to a renewed understanding of the theoretical properties of sampling algorithms [Dal17a; Wib18] building on the seminal work of Jordan, Kinderlehrer and Otto [JKO98].

We consider the following canonical sampling problem. Let π be a log-concave probability measure over $\mathbb{R}^d$ so that π has density equal to e^{-V}, where the potential $V : \mathbb{R}^d \to \mathbb{R}$ is convex. Throughout this paper, we also assume that V is twice continuously differentiable for convenience, though many of our results hold under weaker conditions.

Most MCMC algorithms designed for this problem are based on the *Langevin diffusion* (LD), that is the solution $(X_t)_{t\geq 0}$ to the stochastic differential equation (SDE)

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{LD}$$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

with $(B_t)_{t\geq 0}$ a standard Brownian motion in $\mathbb{R}^d$. Indeed, π is the unique invariant distribution of (LD) and suitable discretizations result in algorithms that can be implemented when V is known only up to an additive constant, which is crucial for applications in Bayesian statistics and machine learning.

A first connection between sampling from log-concave measures and optimizing convex functions is easily seen from (LD): omitting the Brownian motion term yields the gradient flow $\dot{x}_t = -\nabla V(x_t)$, which results in the celebrated gradient descent algorithm when discretized in time [Dal17a; Dal17b]. There is, however, a much deeper connection involving the distribution of X_t rather than X_t itself, and this latter connection has been substantially more fruitful: the marginal distribution of a Langevin diffusion process $(X_t)_{t\geq 0}$ evolves according to a *gradient flow*, over the Wasserstein space of probability measures, that minimizes the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(\cdot \,\|\, \pi)$ [JKO98; AGS08; Vil09]. This point of view has led not only to a better theoretical understanding of the Langevin diffusion [Ber18; CB18; Wib18; DMM19; VW19] but it has also inspired new sampling algorithms based on classical optimization algorithms, such as proximal/splitting methods [Ber18; Wib18; Wib19; SKL20], mirror descent [Hsi+18; Zha+20], Nesterov's accelerated gradient descent [Che+18; Ma+19; DR20], and Newton methods [Mar+12; Sim+16; WL20].

Our contributions. This paper further exploits the optimization perspective on sampling by establishing a theoretical framework for a large class of stochastic processes called *mirror-Langevin diffusions* (MLD) introduced in [Zha+20]. These processes correspond to alternative optimization schemes that minimize the KL divergence over the Wasserstein space by changing its geometry. They show better dependence in key parameters such as the condition number and the dimension.

Our theoretical analysis is streamlined by a technical device which is unexpected at first glance, yet proves to be elegant and effective: we track the progress of these schemes not by measuring the objective function itself, the KL divergence, but rather by measuring the chi-squared divergence to the target distribution π as a surrogate. This perspective highlights the central role of mirror Poincaré inequalities (MP) as sufficient conditions for exponentially fast convergence of the mirror-Langevin diffusion to stationarity in chi-squared divergence, which readily yields convergence in other well-known information divergences, such as the Kullback-Leibler divergence, the Hellinger distance, and the total variation distance [Tsy09, §2.4].

We also specialize our results to the case when the mirror map equals the potential V. This can be understood as the sampling analogue of Newton's method, and we therefore call it the *Newton-Langevin diffusion* (NLD). In this case, the mirror Poincaré inequality translates into the Brascamp-Lieb inequality which automatically holds when V is twice-differentiable and *strictly* convex. In turn, it readily implies exponential convergence of the Newton-Langevin diffusion (Corollary 1) and can be used for approximate sampling even when the second derivative of V vanishes (Corollary 2). Strikingly, the rate of convergence *has no dependence on π or on the dimension d* and, in particular, is robust to cases where $\nabla^2 V$ is arbitrarily close to zero. This *scale-invariant* convergence parallels that of Newton's method in convex optimization and is the first result of this kind for sampling.

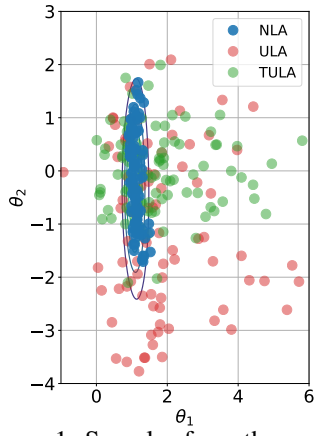

Figure 1: Samples from the posterior distribution of a 2D Bayesian logistic regression model using the Newton-Langevin Algorithm (NLA), the Unadjusted Langevin Algorithm (ULA), and the Tamed Unadjusted Langevin Algorithm (TULA) [Bro+19]. For details, see Section E.2.

This invariance property is useful for approximately sampling from the uniform distribution over a convex body $\mathcal{C}$, which has been well-studied in the computer science literature [FKP94; KLS95; LV07]. By taking the target distribution $\pi \propto \exp(-\beta V)$, where V is any strictly convex *barrier function*, and β, the inverse temperature parameter, is taken to be small (depending on the target accuracy), we can use the Newton-Langevin diffusion, much in the spirit of interior point methods (as promoted by [LTV20]), to output a sample which is approximately uniformly distributed on $\mathcal{C}$; see Corollary 3.

Throughout this paper, we work exclusively in the setting of continuous-time diffusions such as (LD). We refer to the works [DM15; Dal17a; Dal17b; RRT17; CB18; Wib18; DK19; DMM19; DRK19; Mou+19; VW19] for discretization error bounds, and leave this question open for future works.

Related work. The discretized Langevin algorithm, and the Metropolis-Hastings adjusted version, have been well-studied when used to sample from strongly log-concave distributions, or distributions satisfying a log-Sobolev inequality [Dal17b; DM17; CB18; Che+19; DK19; DM+19; Dwi+19; Mou+19; VW19]. Moreover, various ways of adapting Langevin diffusion to sample from bounded domains have been proposed [BEL18; Hsi+18; Zha+20]; in particular, [Zha+20] studied the discretized mirror-Langevin diffusion. Finally, we note that while our analysis and methods are inspired by the optimization perspective on sampling, it connects to a more traditional analysis based on coupling stochastic processes. Quantitative analysis of the continuous Langevin diffusion process associated to SDE (LD) has been performed with Poincaré and log-Sobolev inequalities [BGG12; BGL14; VW19], and with couplings of stochastic processes [CL89; Ebe16].

Notation. The Euclidean norm over $\mathbb{R}^d$ is denoted by $\|\cdot\|$. Throughout, we simply write $\int g$ to denote the integral with respect to the Lebesgue measure: $\int g(x)\,\mathrm{d}x$. When the integral is with respect to a different measure μ, we explicitly write $\int g\,\mathrm{d}\mu$. The expectation and variance of $g(X)$ when $X \sim \mu$ are respectively denoted $\mathbb{E}_\mu\, g = \int g\,\mathrm{d}\mu$ and $\mathrm{var}_\mu\, g := \int (g - \mathbb{E}_\mu\, g)^2\,\mathrm{d}\mu$. When clear from context, we sometimes abuse notation by identifying a measure μ with its Lebesgue density.

2 Mirror-Langevin diffusions

Before introducing mirror-Langevin diffusions, our main objects of interest, we provide some intuition for their construction by drawing a parallel with convex optimization.

2.1 Gradient flows, mirror flows, and Newton's method

We briefly recall some background on gradient flows and mirror flows; we refer readers to the monograph [Bub15] for the convergence analysis of the corresponding discrete-time algorithms.

Suppose we want to minimize a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. The *gradient flow* of f is the curve $(x_t)_{t\geq 0}$ on $\mathbb{R}^d$ solving $\dot{x}_t = -\nabla f(x_t)$. A suitable time discretization of this curve yields the well-known *gradient descent* (GD).

Although the gradient flow typically works well for optimization over Euclidean spaces, it may suffer from poor dimension scaling in more general cases such as Banach space optimization; a notable example is the case when f is defined over the probability simplex equipped with the ℓ_1 norm. This observation led Nemirovskii and Yudin [NJ79] to introduce the *mirror flow*, which is defined as follows. Let $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a *mirror map*, that is a strictly convex twice continuously differentiable function of *Legendre type*[1]. The mirror flow $(x_t)_{t\geq 0}$ satisfies $\partial_t \nabla\phi(x_t) = -\nabla f(x_t)$, or equivalently, $\dot{x}_t = -[\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)$. The corresponding discrete-time algorithms, called *mirror descent* (MD) algorithms, have been successfully employed in varied tasks of machine learning [Bub15] and online optimization [BC12] where the entropic mirror map plays an important role. In this work, we are primarily concerned with the following choices for the mirror map:

1. When $\phi = \|\cdot\|^2/2$, then the mirror flow reduces to the gradient flow.
2. Taking $\phi = f$ and the discretization $x_{k+1} = x_k - h_k\,[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$ yields another popular optimization algorithm known as (damped) *Newton's method*. Newton's method has the important property of being invariant under affine transformations of the problem, and its local convergence is known to be much faster than that of GD; see [Bub15, §5.3].

2.2 Mirror-Langevin diffusions

We now introduce the *mirror-Langevin diffusion* (MLD) of [Zha+20]. Just as LD corresponds to the gradient flow, the MLD is the sampling analogue of the mirror flow. To describe it, let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a mirror map as in the previous section. Then, the mirror-Langevin diffusion satisfies the SDE

$$X_t = \nabla\phi^\star(Y_t), \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2\phi(X_t)]^{1/2}\,\mathrm{d}B_t\,, \tag{MLD}$$

[1] This ensures that $\nabla\phi$ is invertible, c.f. [Roc97, §26].

where $\phi^\star$ denotes the convex conjugate of ϕ [BL06, §3.3]. In particular, if we choose the mirror map ϕ to equal the potential V, then we arrive at a sampling analogue of *Newton's method*, which we call the *Newton-Langevin diffusion* (NLD),

$$X_t = \nabla V^\star(Y_t), \qquad \mathrm{d}Y_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,[\nabla^2 V(X_t)]^{1/2}\,\mathrm{d}B_t. \tag{NLD}$$

From our intuition gained from optimization, we expect that NLD has special properties, such as affine invariance and faster convergence. We validate this intuition in Corollary 1 below by showing that, provided π is strictly log-concave, the NLD converges to stationarity exponentially fast, with no dependence on π. This should be contrasted with the vanilla Langevin diffusion (LD), for which the convergence rate depends on the Poincaré constant of π, as we discuss in the next section.

We end this section by comparing MLD and NLD with similar sampling algorithms proposed in the literature inspired by mirror descent and Newton's method.

Mirrored Langevin dynamics. A variant of MLD, called "mirrored Langevin dynamics", was introduced in [Hsi+18]. The mirrored Langevin dynamics is motivated by constrained sampling and corresponds to the vanilla Langevin algorithm applied to the new target measure $(\nabla\phi)_{\#}\pi$. In contrast, MLD can be understood as a Riemannian diffusion w.r.t. the Riemannian metric induced by the mirror map ϕ. Thus, the motivations and properties of the two algorithms are different, and we refer to [Zha+20] for further comparison of the two algorithms.

An earlier draft of [Hsi+18] also introduced MLD, along with a continuous-time analysis of the diffusion. Their convergence analysis is based on the classical Bakry-Émery criterion (see [BGL14]), which is generally harder to check than the mirror Poincaré inequality (MP) that we introduce below; in particular, when $\phi = V$, we show that the mirror Poincaré inequality holds automatically.

Quasi-Newton diffusion. The paper [Sim+16] proposes a quasi-Newton sampling algorithm, based on L-BFGS, which is partly motivated by the desire to avoid computation of the third derivative $\nabla^3 V$ while implementing the Newton-Langevin diffusion. We remark, however, that the form of NLD employed above, which treats V as a mirror map, does not in fact require the computation of $\nabla^3 V$, and thus can be implemented practically; see Section 5. Moreover, since we analyze the full NLD, rather than a quasi-Newton implementation, we are able to give a clean convergence result.

Information Newton's flow. Inspired by the perspective of [JKO98], which views the Langevin diffusion as a gradient flow in the Wasserstein space of probability measures, the paper [WL20] proposes an approach termed "information Newton's flow" that applies Newton's method directly on the space of probability measures equipped with either the Fisher-Rao or the Wasserstein metric. However, unlike LD and NLD that both operate at the level of particles, information Newton's flow faces significant challenges at the level of both implementation and analysis.

3 Convergence analysis

3.1 Convergence of gradient flows and mirror flows

We provide a brief reminder about the convergence analysis of gradient flows and mirror flows defined in Section 2.1 to provide intuition for the next section. Throughout, let f be a differentiable function with minimizer x^*.

Consider first the gradient flow for f: $\dot{x}_t = -\nabla f(x_t)$. We get $\partial_t[f(x_t) - f(x^*)] = -\|\nabla f(x_t)\|^2$ from a straightforward computation. From this identity, it is natural to assume a *Polyak-Łojasiewicz* (PL) *inequality*, which is well-known in the optimization literature [KNS16] and can be employed even when f is not convex [Che+20]. Indeed, if there exists a constant $C_{\mathsf{PL}} > 0$ with

$$f(x) - f(x^*) \le \frac{C_{\mathsf{PL}}}{2}\,\|\nabla f(x)\|^2 \qquad \forall\, x \in \mathbb{R}^d\,, \tag{PL}$$

then $\partial_t[f(x_t) - f(x^*)] \le -\frac{2}{C_{\mathsf{PL}}}\,[f(x_t) - f(x^*)]$. Together with Grönwall's inequality, it readily yields exponentially fast convergence in objective value: $f(x_t) \le f(x_0)\,e^{-2t/C_{\mathsf{PL}}}$.

A similar analysis may be carried out for the mirror flow. Fix a mirror map ϕ and consider the mirror flow: $\dot{x}_t = -[\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)$. It holds $\partial_t[f(x_t) - f(x^*)] = -\langle \nabla f(x_t), [\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)\rangle$.

Therefore, the analogue of (PL) which guarantees exponential decay in the objective value is the following inequality, which we call a *mirror PL inequality*:

$$f(x) - f(x^*) \le \frac{C_{\mathsf{MPL}}}{2} \langle \nabla f(x), [\nabla^2 \phi(x)]^{-1} \nabla f(x) \rangle \qquad \forall x \in \mathbb{R}^d. \tag{MPL}$$

Next, we describe analogues of (PL) and (MPL) that guarantee convergence of LD and MLD.

3.2 Convergence of mirror-Langevin diffusions

The above analysis employs the objective function f to measure the progress of both the gradient and mirror flows. While this is the most natural choice, our approach below crucially relies on measuring progress via a *different functional* F. What should we use as F? To answer this question, we first consider the simpler case of the vanilla Langevin diffusion (LD), which is a special case of MLD when the mirror map is $\phi = \|\cdot\|^2/2$. We keep this discussion informal and postpone rigorous arguments to Appendix A.

Since the work of [JKO98], it has been known that the marginal distribution μ_t at time $t \ge 0$ of LD evolves according to the *gradient flow* of the KL divergence $D_{\mathrm{KL}}(\cdot \parallel \pi)$ with respect to the 2-Wasserstein distance W_2; we refer readers to [San17] for an overview of this work, and to [AGS08; Vil09] for comprehensive treatments. Therefore, the most natural choice for F is, as in Section 3.1, the objective function $D_{\mathrm{KL}}(\cdot \parallel \pi)$ itself. Following this approach, one can compute [Vil03, §9.1.5]

$$\partial_t D_{\mathrm{KL}}(\mu_t \parallel \pi) = -\int \Big\| \nabla \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \Big\|^2 \mathrm{d}\mu_t = -4 \int \Big\| \nabla \sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}} \Big\|^2 \mathrm{d}\pi.$$

In this setup, the role of the PL inequality (PL) is played by a *log-Sobolev inequality* of the form

$$\mathrm{ent}_\pi(g^2) := \int g^2 \ln(g^2) \,\mathrm{d}\pi - \Big(\int g^2 \,\mathrm{d}\pi \Big) \ln \Big(\int g^2 \,\mathrm{d}\pi \Big) \le 2 C_{\mathsf{LSI}} \int \|\nabla g\|^2 \,\mathrm{d}\pi. \tag{LSI}$$

When $g = \sqrt{\mathrm{d}\mu_t/\mathrm{d}\pi}$, (LSI) reads $D_{\mathrm{KL}}(\mu_t \parallel \pi) \le 2 C_{\mathsf{LSI}} \int \big\| \nabla \sqrt{\mathrm{d}\mu_t/\mathrm{d}\pi} \big\|^2 \mathrm{d}\pi$, which implies exponentially fast convergence: $D_{\mathrm{KL}}(\mu_t \parallel \pi) \le D_{\mathrm{KL}}(\mu_0 \parallel \pi)\, e^{-2t/C_{\mathsf{LSI}}}$ by Grönwall's inequality.

A disadvantage of this approach, however, is that the log-Sobolev inequality (LSI) does not hold for any log-concave measure π, or it may hold with a poor constant C_{LSI}. For example, it is known that the log-Sobolev constant of an isotropic log-concave distribution must in general depend on the diameter of its support [LV18]. In contrast, we work below with a *Poincaré inequality*, which is conjecturally satisfied by such distributions with a *universal constant* [KLS95].

Motivated by [BCG08; CG09], we instead consider the *chi-squared divergence*

$$F(\mu) = \chi^2(\mu \parallel \pi) := \mathrm{var}_\pi \frac{\mathrm{d}\mu}{\mathrm{d}\pi} = \int \Big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \Big)^2 \mathrm{d}\pi - 1, \qquad \text{if } \mu \ll \pi\,,$$

and $F(\mu) = \infty$ otherwise. It is well-known that the law $(\mu_t)_{t \ge 0}$ of LD satisfies the Fokker-Planck equation in the weak sense [KS91, §5.7]:

$$\partial_t \mu_t = \mathrm{div}\big(\mu_t \nabla \ln \frac{\mu_t}{\pi} \big).$$

Using this, we can compute the derivative of the chi-squared divergence:

$$\frac{1}{2} \partial_t F(\mu_t) = \int \frac{\mu_t}{\pi}\, \partial_t \mu_t = \int \frac{\mu_t}{\pi}\, \mathrm{div}\big(\mu_t \nabla \ln \frac{\mu_t}{\pi} \big) = -\int \big\langle \nabla \ln \frac{\mu_t}{\pi}, \nabla \frac{\mu_t}{\pi} \big\rangle\, \mu_t = -\int \Big\| \nabla \frac{\mu_t}{\pi} \Big\|^2 \pi\,,$$

and exponential convergence of the chi-squared divergence follows if π satisfies a Poincaré inequality:

$$\mathrm{var}_\pi\, g \le C_{\mathsf{P}}\, \mathbb{E}_\pi[\|\nabla g\|^2] \qquad \text{for all locally Lipschitz } g \in L^2(\pi). \tag{P}$$

Thus, when using the chi-squared divergence to track progress, the role of the PL inequality is played by a Poincaré inequality. As we discuss in Sections 4.1 and 4.3 below, the Poincaré inequality is significantly weaker than the log-Sobolev inequality.

A similar analysis may be carried out for MLD using an appropriate variation of Poincaré inequalities.

Definition 1 (Mirror Poincaré inequality)**.** Given a mirror map ϕ, we say that the distribution π satisfies a *mirror Poincaré inequality* with constant C_{MP} if

$$\operatorname{var}_\pi g \le C_{\mathsf{MP}}\, \mathbb{E}_\pi \langle \nabla g, (\nabla^2 \phi)^{-1} \nabla g \rangle \qquad \text{for all locally Lipschitz } g \in L^2(\pi). \tag{MP}$$

When $\phi = \|\cdot\|^2/2$, (MP) is simply called a *Poincaré inequality* and the smallest C_{MP} for which the inequality holds is the *Poincaré constant* of π, denoted C_{P}.

Using a similar argument as the one above, we show exponential convergence of MLD in $\chi^2(\cdot \,\|\, \pi)$ under (MP). Together with standard comparison inequalities between information divergences [Tsy09, §2.4], it implies exponential convergence in a variety of commonly used divergences, including the total variation (TV) distance $\|\cdot - \pi\|_{\mathrm{TV}}$, the Hellinger distance $H(\cdot, \pi)$, and the KL divergence $D_{\mathrm{KL}}(\cdot \,\|\, \pi)$.

Theorem 1. *For each $t \ge 0$, let μ_t be the marginal distribution of MLD with target distribution π at time t. Then if π satisfies the mirror Poincaré inequality (MP) with constant C_{MP}, it holds*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \pi) \le e^{-\frac{2t}{C_{\mathsf{MP}}}} \chi^2(\mu_0 \,\|\, \pi), \quad \forall\, t \ge 0\,.$$

We give two proofs of this result in Appendix A.

Recall that LD can be understood as a gradient flow for the KL divergence on the 2-Wasserstein space. In light of this interpretation, the above bound for the KL divergence yields a convergence rate *in objective value*, and it is natural to wonder whether a similar rate holds for the iterates themselves in terms of 2-Wasserstein distance. From the works [Din15; Led18; Liu20], it is known that a Poincaré inequality (P) implies the transportation-cost inequality

$$W_2^2(\mu, \pi) \le 2C_{\mathsf{P}} \chi^2(\mu \,\|\, \pi), \qquad \forall \mu \ll \pi. \tag{1}$$

Initially unaware of these works, we proved that a Poincaré inequality implies a suboptimal chi-squared transportation inequality. Since the suboptimal inequality already suffices for our purposes, we state and prove it in Appendix B. We thank Jon Niles-Weed for bringing this to our attention.

The inequality (1) implies that if π has a finite Poincaré constant C_{P} then Theorem 1 also yields exponential convergence in Wasserstein distance. In the rest of the paper, we write this result as

$$\frac{1}{2C_{\mathsf{P}}} W_2^2(\mu_t, \pi) \le e^{-\frac{2t}{C_{\mathsf{MP}}}} \chi^2(\mu_0 \,\|\, \pi)\,,$$

for *any* target measure π that satisfies a mirror Poincaré inequality, with the convention that $C_{\mathsf{P}} = \infty$ when π fails to satisfy a Poincaré inequality. In this case, the above inequality is simply vacuous.

4 Applications

We specialize Theorem 1 to the following important applications.

4.1 Newton-Langevin diffusion

For NLD, we assume that V is strictly convex and twice continuously differentiable; take $\phi = V$. In this case, the mirror Poincaré inequality (MP) reduces to the *Brascamp-Lieb inequality*, which is known to hold with constant $C_{\mathsf{MP}} = 1$ for any strictly log-concave distribution π [BL76; BL00; Gen08]. It yields the following remarkable result where the exponential contraction rate has no dependence on π nor on the dimension d.

Corollary 1. *Suppose that V is strictly convex and twice continuously differentiable. Then, the law $(\mu_t)_{t\ge 0}$ of NLD satisfies*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \mu),\ \frac{1}{2C_{\mathsf{P}}} W_2^2(\mu_t, \pi) \le e^{-2t} \chi^2(\mu_0 \,\|\, \pi).$$

If π is log-concave, then it satisfies a Poincaré inequality [AB15; LV17] so that the result in Wasserstein distance holds. In fact, contingent on the famous *Kannan-Lovász-Simonovitz* (KLS) conjecture ([KLS95]), the Poincaré constant of any log-concave distribution π is upper bounded by a constant, independent of the dimension, times the largest eigenvalue of the covariance matrix of π.

At this point, one may wonder, under the same assumptions as the Brascamp-Lieb inequality, whether a mirror version of the log-Sobolev inequality (LSI) holds. This question was answered negatively in [BL00], thus reinforcing our use of the chi-squared divergence as a surrogate for the KL divergence.

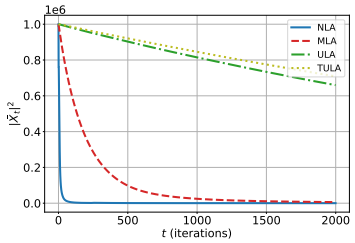

Figure 2: Approximately sampling from $\pi \propto e^{-\|\cdot\|}$ by sampling from $\pi_\beta \propto e^{-\|\cdot\|-\beta\|\cdot-\mathbf{1}\|^2}$ ($\beta = .0005$). Algorithms are initialized at a random X_0 with $\|X_0\| = 1000$. The plot shows the squared distance of the running means to 0.

If the potential V is convex, but degenerate (i.e., not strictly convex) we cannot use NLD directly with π as the target distribution. Instead, we perturb π slightly to a new measure π_β, which is strongly log-concave, and for which we can use NLD. Crucially, due to the scale invariance of NLD, the time it takes for NLD to mix does not depend on β, the parameter which governs the approximation error.

Corollary 2. *Fix a target accuracy $\varepsilon > 0$. Suppose $\pi = e^{-V}$ is log-concave and set $\pi_\beta \propto e^{-V-\beta\|\cdot\|^2}$, where $\beta \le \varepsilon^2/(2\int \|\cdot\|^2 \,\mathrm{d}\pi)$. Then, the law $(\mu_t)_{t\ge 0}$ of NLD with target distribution π_β satisfies $\|\mu_t - \pi\|_{\mathrm{TV}} \le \varepsilon$ by time $t = \frac{1}{2}\ln[2\chi^2(\mu_0 \parallel \pi_\beta)] + \ln(1/\varepsilon)$.*

Proof. From our assumption, it holds

$$D_{\mathrm{KL}}(\pi \parallel \pi_\beta) = \int \ln \frac{\mathrm{d}\pi}{\mathrm{d}\pi_\beta}\,\mathrm{d}\pi = \beta \int \|\cdot\|^2 \,\mathrm{d}\pi + \ln \int e^{-\beta\|\cdot\|^2}\,\mathrm{d}\pi \le \beta \int \|\cdot\|^2 \,\mathrm{d}\pi \le \frac{\varepsilon^2}{2}.$$

Moreover, Theorem 1 with the above choice of t yields $D_{\mathrm{KL}}(\mu_t \parallel \pi_\beta) \le \varepsilon^2/2$. To conclude, we use Pinsker's inequality and the triangle inequality for $\|\cdot\|_{\mathrm{TV}}$. □

Convergence guarantees for other cases where ϕ is only a *proxy* for V are presented in Appendix C.

4.2 Sampling from the uniform distribution on a convex body

Next, we consider an application of NLD to the problem of sampling from the uniform distribution π on a convex body $\mathcal{C}$. A natural method of outputting an approximate sample from π is to take a strictly convex function $\widetilde{V} : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that $\mathrm{dom}\,\widetilde{V} = \mathcal{C}$ and $\widetilde{V}(x) \to \infty$ as $x \to \partial\mathcal{C}$, and to run NLD with target distribution $\pi_\beta \propto e^{-\beta\widetilde{V}}$, where the inverse temperature β is taken to be small (so that $\pi_\beta \approx \pi$). The function $\widetilde{V}$ is known as a *barrier function*.

Although we can take any choice of barrier function $\widetilde{V}$, we obtain a clean theoretical result if we assume that $\widetilde{V}$ is ν^{-1}-exp-concave, that is, the mapping $\exp(-\nu^{-1}\widetilde{V})$ is concave. Interestingly, this assumption further deepens the rich analogy between sampling and optimization, since such barriers are widely studied in the optimization literature. There, the property of exp-concavity is typically paired with the property of *self-concordance*, and barrier functions satisfying these two properties are a cornerstone of the theory of *interior point algorithms* (see [Bub15, §5.3] and [Nes04, §4]).

We now formulate our sampling result. In our continuous framework, it does not require self-concordance of the barrier function.

Corollary 3. *Fix a target accuracy $\varepsilon > 0$. Let π be the uniform distribution over a convex body $\mathcal{C}$ and let $\widetilde{V}$ be a ν^{-1}-exp-concave barrier for $\mathcal{C}$. Then, the law $(\mu_t)_{t\ge 0}$ of NLD with target density $\pi_\beta \propto e^{-\beta\widetilde{V}}$ for $\beta \le \varepsilon^2/(2\nu)$ satisfies $\|\mu_t - \pi\|_{\mathrm{TV}} \le \varepsilon$ by time $t = \frac{1}{2}\ln[2\chi^2(\mu_0 \parallel \pi_\beta)] + \ln(1/\varepsilon)$.*

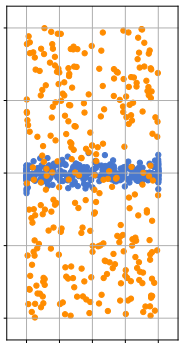

Figure 3: Uniform sampling from the set $[-0.01, 0.01] \times [-1, 1]$: PLA (blue) vs. NLA (orange). See Section E.3.

Proof. Lemma 1 in Appendix D ensures that $D_{\mathrm{KL}}(\pi_\beta \,\|\, \pi) \le \varepsilon^2/2$. We conclude as in the proof of Corollary 2, by using Theorem 1, Pinsker's inequality, and the triangle inequality for $\|\cdot\|_{\mathrm{TV}}$. □

We demonstrate the efficacy of NLD in a simple simulation: sampling uniformly from the ill-conditioned rectangle $[-a, a] \times [-1, 1]$ with $a = 0.01$ (Figure 3). We compare NLA with the Projected Langevin Algorithm (PLA) [BEL18], both with 200 iterations and $h = 10^{-4}$. For NLA, we take $\widetilde{V}(x) = -\log(1 - x_1^2) - \log(a^2 - x_2^2)$ and $\beta = 10^{-4}$.

4.3 Langevin diffusion under a Poincaré inequality

We conclude this section by giving some implications of Theorem 1 to the classical Langevin diffusion (LD) when $\phi = \|\cdot\|^2/2$. In this case, the mirror Poincaré inequality (MP) reduces to the classical Poincaré inequality (P) as in Section 3.2.

Corollary 4. *Suppose that π satisfies a Poincaré inequality (P) with constant $C_{\mathsf{P}} > 0$. Then, the law $(\mu_t)_{t\ge 0}$ of the Langevin diffusion (LD) satisfies*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \mu),\ \frac{1}{2C_{\mathsf{P}}} W_2^2(\mu_t, \pi) \le e^{-\frac{2t}{C_{\mathsf{P}}}} \chi^2(\mu_0 \,\|\, \pi).$$

The convergence in TV distance recovers results of [Dal17b; DM17]. Bounds for the stronger error metric $\chi^2(\cdot \,\|\, \pi)$ have appeared explicitly in [CLL19; VW19] and is implicit in the work of [BCG08; CG09] on which the TV bound of [DM17] is based.

Moreover, it is classical that if π satisfies a log-Sobolev inequality (LSI) with constant C_{LSI} then it has Poincaré constant $C_{\mathsf{P}} \le C_{\mathsf{LSI}}$. Thus, the choice of the chi-squared divergence as a surrogate for the KL divergence when tracking progress indeed requires weaker assumptions on π.

5 Numerical experiments

In this section, we examine the numerical performance of the *Newton-Langevin Algorithm* (NLA), which is given by the following Euler discretization of NLD:

$$\nabla V(X_{k+1}) = (1 - h)\nabla V(X_k) + \sqrt{2h}\,[\nabla^2 V(X_k)]^{1/2}\xi_k, \tag{NLA}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. $\mathcal{N}(0, I_d)$ variables. In cases where ∇V does not have a closed-form inverse, such as the logistic regression case of Section E.2, we invert it numerically by solving the convex optimization problem $\nabla V^\star(y) = \mathrm{argmax}_{x\in\mathbb{R}^d}\{\langle x, y\rangle - V(x)\}$.

We focus here on sampling from an ill-conditioned generalized Gaussian distribution on $\mathbb{R}^{100}$ with $V(x) = \langle x, \Sigma^{-1}x\rangle^\gamma/2$ for $\gamma = 3/4$ to demonstrate the scale invariance of NLD established in Corollary 1. Additional experiments, including the Gaussian case $\gamma = 1$, are given in Appendix E.

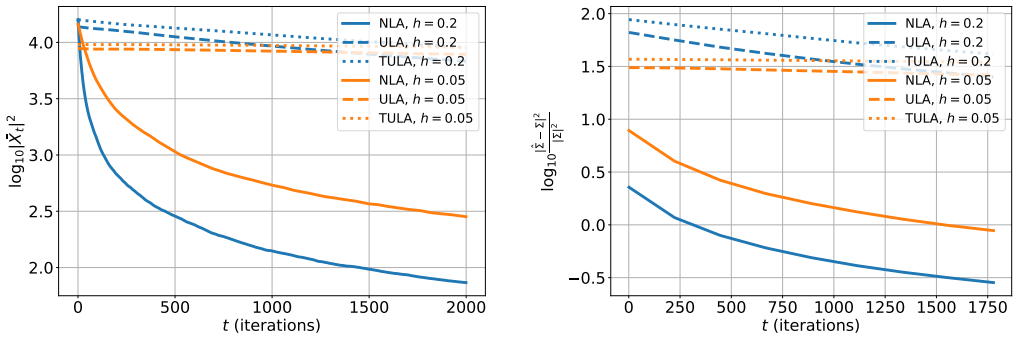

Figure 4: $V(x) = \langle x, \Sigma^{-1}x\rangle^{3/4}/2$, $\Sigma = \mathrm{diag}(1, 2, \ldots, 100)$. Left: absolute squared error of the mean 0. Right: relative squared error for the scatter matrix Σ.

Figure 4 compares the performance of NLA to that of the Unadjusted Langevin Algorithm (ULA) [DM+19] and of the Tamed Unadjusted Langevin Algorithm (TULA) [Bro+19]. We run the

algorithms 50 times and compute running estimates for the mean and scatter matrix of the family following [ZWG13]. Convergence is measured in terms of squared distance between means and relative squared distance between scatter matrices, $\|\hat{\Sigma} - \Sigma\|^2/\|\Sigma\|^2$. NLA generates samples that rapidly approximate the true distribution and also displays stability to the choice of the step size h.

6 Open questions

We conclude this paper by discussing several intriguing directions for future research. In this paper, we focused on giving clean convergence results for the continuous-time diffusions MLD and NLD, and we leave open the problem of obtaining discretization error bounds. In discrete time, Newton's method can be unstable, and one uses methods such as damped Newton, Levenburg-Marquardt, or cubic-regularized Newton [CGT00; NP06]; it is an interesting question to develop sampling analogues of these optimization methods. In a different direction, we ask the following question: are there appropriate variants of other popular sampling methods, such as accelerated Langevin [Ma+19] or Hamiltonian Monte Carlo [Nea12], which also enjoy the scale invariance of NLD?

A Proof of Theorem 1

The law $(\mu_t)_{t\geq 0}$ of MLD satisfies the Fokker-Planck equation

$$\partial_t \mu_t = \operatorname{div}\big(\mu_t \, (\nabla^2 \phi)^{-1} \nabla \ln \frac{\mu_t}{\pi}\big). \tag{2}$$

A unique solution to this equation, with enough regularity to justify our computations below, exists under fairly benign conditions on ϕ and V, see [LL08, Proposition 6].

As discussed in Section 3.2, it suffices to prove the convergence result in chi-squared divergence. The convergence results for total variation distance, Hellinger distance, and KL divergence follow from the inequalities [Tsy09, §2.4]

$$2\|\mu - \pi\|_{\mathrm{TV}}^2, \; H^2(\mu,\pi), \; D_{\mathrm{KL}}(\mu \parallel \pi) \leq \chi^2(\mu \parallel \pi), \qquad \forall \mu \ll \pi,$$

while the convergence in Wasserstein distance follows from (1).

Proof of Theorem 1. Using the Fokker-Planck equation (2), we may compute

$$\begin{aligned}\partial_t \chi^2(\mu_t \parallel \pi) &= \partial_t \int \frac{\mu_t^2}{\pi} = 2\int \frac{\mu_t}{\pi}\, \partial_t \mu_t = 2\int \frac{\mu_t}{\pi}\, \operatorname{div}\big(\mu_t \, (\nabla^2\phi)^{-1}\nabla \ln \frac{\mu_t}{\pi}\big) \\ &= -2\int \big\langle \nabla \frac{\mu_t}{\pi}, (\nabla^2\phi)^{-1}\nabla \ln \frac{\mu_t}{\pi}\big\rangle \, \mu_t = -2\int \big\langle \nabla \frac{\mu_t}{\pi}, (\nabla^2\phi)^{-1}\nabla \frac{\mu_t}{\pi}\big\rangle \, \pi.\end{aligned}$$

The mirror Poincaré inequality (MP) implies that this quantity is at most $-2C_{\mathsf{MP}}^{-1}\chi^2(\mu_t \parallel \pi)$, which completes the proof via Grönwall's inequality. □

We may reinterpret this proof within Markov semigroup theory.

Proof of Theorem 1 from a Markov semigroup perspective. We denote by $(P_t)_{t\geq 0}$ the semigroup of MLD; we refer readers to [BGL14; Han16] for background on Markov semigroup theory. The Dirichlet form $\mathcal{E}$ is given by

$$\mathcal{E}(f,g) = \int \langle \nabla f, (\nabla^2\phi)^{-1}\nabla g\rangle \, \mathrm{d}\pi.$$

Since it is a self-adjoint semigroup, we get for all $f \in L^2(\pi)$,

$$\int P_t(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}) f \, \mathrm{d}\pi = \int (\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}) P_t f \, \mathrm{d}\pi = \int P_t f \, \mathrm{d}\mu_0 = \int f \, \mathrm{d}\mu_t = \int \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} f \, \mathrm{d}\pi \;,$$

so that

$$P_t(\frac{\mu_0}{\pi}) = \frac{\mu_t}{\pi}.$$

Therefore,

$$\chi^2(\mu_t \,\|\, \pi) := \mathrm{var}_\pi\big(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\big) = \mathrm{var}_\pi\, P_t\big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big).$$

Using a classical result of Markov semigroup theory (see for instance [CG09, Theorem 2.1] or [BGL14, Theorem 4.2.5])

$$\chi^2(\mu_t \,\|\, \pi) = \mathrm{var}_\pi\, P_t\big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big) \le e^{-\frac{2t}{C}}\, \mathrm{var}_\pi\big(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\big) = e^{-\frac{2t}{C}} \chi^2(\mu_0 \,\|\, \pi)$$

if and only if the semigroup $(P_t)_{t\ge 0}$ satisfies

$$\mathrm{var}_\pi(f) \le C\mathcal{E}(g,g), \qquad \text{for all } \ g \in D(\mathcal{E}), \tag{3}$$

where $\mathcal{E}$ is the Dirichlet form of $(P_t)_{t\ge 0}$ with domain $D(\mathcal{E})$. To conclude the proof, it suffices to note that (3) is precisely our assumption (MP) with $C = C_{\mathsf{MP}}$. □

B Convergence in 2-Wasserstein distance

B.1 Background

As we have discussed, the proof of Theorem 1 in Appendix A implies that for any strictly log-concave target measure, the Newton-Langevin diffusion converges exponentially fast in the following error metrics: chi-squared divergence, KL divergence, Hellinger distance, and total variation distance. We also remark that convergence in Rényi divergences can also be proved in this setting, as in [VW19]. On the other hand, we would also like to know if we can obtain convergence results for *optimal transport* distances [Vil03]. As a first step, the transportation inequality of [Cor17],

$$D_{\mathrm{KL}}(\mu \,\|\, \pi) \ge \mathcal{T}_{D_V}(\mu \,\|\, \pi) := \inf\{\mathbb{E}\, D_V(X \,\|\, Z) : (X,Z) \text{ is a coupling of } (\mu,\pi)\},$$

which holds for all $\mu \ll \pi$, implies exponentially fast convergence in the asymmetric transportation cost $\mathcal{T}_{D_V}$, where $D_V(\cdot \,\|\, \cdot)$ is the Bregman divergence associated with V.

We turn towards the question of convergence in the 2-Wasserstein distance (denoted W_2). When π is strongly log-concave, there is an elegant direct proof of exponential contraction in W_2 via a coupling of the Langevin process (see [Vil03, Exercise 9.10]). In general, however, convergence in W_2 is typically deduced from convergence in KL divergence, with the help of a *transportation-cost inequality*

$$W_2^2(\mu,\pi) \le C D_{\mathrm{KL}}(\mu \,\|\, \pi). \tag{4}$$

It has been known since the work of [OV00] that a log-Sobolev inequality (LSI) with constant C_{LSI} implies the validity of (4) with constant $C = C_{\mathsf{LSI}}$. Since an LSI may not always hold or may hold with a poor constant, [BV05] provides weaker conditions: namely, if there exists $\alpha > 0$ such that

$$\int \exp(\alpha \|x - x_0\|^2)\, \mathrm{d}\pi(x) < \infty, \tag{5}$$

then we have the weaker inequality

$$W_2^2(\mu,\pi) \lesssim D_{\mathrm{KL}}(\mu \,\|\, \pi) + \sqrt{D_{\mathrm{KL}}(\mu \,\|\, \pi)}.$$

Therefore, either the validity of an LSI or a square exponential moment suffice to transfer convergence in KL divergence to convergence in W_2. In fact, it turns out that the log-Sobolev inequality (LSI), the transportation inequality (4), and the square exponential moment condition (5) are all equivalent for log-concave measures,and they are in general strictly stronger than the Poincaré inequality (P) [Bob99; OV00; BV05].

Since Theorem 1 provides a stronger control, namely in chi-squared divergence rather than in KL divergence, the reader might wonder if a weaker transportation inequality in which the RHS of (4) is replaced by $C\chi^2(\mu \,\|\, \pi)^{1/p}$ might hold under weaker assumptions. Indeed, the recent works [Din15; Led18; Liu20] answer this question positively by showing that the Poincaré inequality (P) implies the transportation-cost inequality

$$W_2^2(\mu,\nu) \le C \inf_{p\ge 1}\{p^2 \chi^2(\mu \,\|\, \pi)^{1/p}\}, \qquad \forall \mu \ll \pi \tag{6}$$

with constant $C = 2C_{\mathsf{P}}$. In fact, the converse also holds: the validity of (6) implies the Poincaré inequality (P) with constant $C_{\mathsf{P}} = C/\sqrt{2}$.

If we specialize this result to the case $p = 2$, then the Poincaré inequality (P) implies

$$W_2^2(\mu, \nu) \le 8C_{\mathsf{P}} \sqrt{\chi^2(\mu \,\|\, \pi)}, \qquad \forall \mu \ll \pi. \tag{7}$$

In the next section, we give a proof of the inequality (7) with a slightly worse constant, i.e., with 9 instead of 8.

We now briefly describe the method of [OV00], since it is relevant for our approach. Otto and Villani work in the framework of *Otto calculus*, which interprets LD as the gradient flow of the KL divergence in the space of probability measures equipped with the W_2 metric. As discussed in Section 3.2, an LSI is a PL inequality, which ensures rapid convergence of the gradient flow. This is then used to deduce the transportation-cost inequality (4).

We follow the argument of Otto and Villani, but consider the *gradient flow of the chi-squared divergence* instead of the KL divergence. We prove a Łojasiewicz inequality for the chi-squared divergence, and use the gradient flow to deduce (7) (with a slightly worse constant).

B.2 Proof of the chi-squared transportation inequality

Following the proof outline above, we start by proving a PL-type inequality for the chi-squared divergence. Using tools developed in [AGS08], it is a standard exercise to establish that the Wasserstein gradient of the functional $\mu \mapsto \chi^2(\mu \,\|\, \pi)$ is given by $2\nabla(\mathrm{d}\mu/\mathrm{d}\pi)$. Therefore, the right-hand side of the following inequality involves the squared norm of the Wasserstein gradient of the chi-squared divergence, where we use the norm corresponding to the Riemannian structure of Wasserstein space (see [AGS08, §8]). Note that since the objective is raised to the power $3/2$ on the left-hand side it is not quite a PL inequality, and rather it is a form commonly referred to as a Łojasiewicz inequality [Loj63] with parameter $3/4$.

Proposition 1. *Let $C_{\mathsf{P}} \in (0, \infty]$ denote the Poincaré constant of π. Then,*

$$\chi^2(\mu \,\|\, \pi)^{3/2} \le \frac{9C_{\mathsf{P}}}{4} \int \|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^2 \,\mathrm{d}\mu, \qquad \forall \mu \ll \pi.$$

Proof. Using the Poincaré inequality (P), we obtain

$$\int \|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^2 \,\mathrm{d}\mu = \int \|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^2 \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \,\mathrm{d}\pi = \frac{4}{9} \int \|\nabla (\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2}\|^2 \,\mathrm{d}\pi \ge \frac{4}{9C_{\mathsf{P}}} \operatorname{var}_\pi\big((\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2}\big).$$

In the following steps, we apply the following: (1) $\operatorname{var} X \le \mathbb{E}[|X - c|^2]$ for any $c \in \mathbb{R}$; (2) $x \mapsto x^{2/3}$ is $2/3$-Hölder continuous with unit constant; (3) Jensen's inequality.

$$\begin{aligned}
\chi^2(\mu \,\|\, \pi) = \operatorname{var}_\pi\big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\big) &\overset{(1)}{\le} \mathbb{E}_\pi\Big[\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - \mathbb{E}_\pi\big[(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2}\big]^{2/3}\Big|^2\Big] \\
&\overset{(2)}{\le} \mathbb{E}_\pi\Big[\Big|(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2} - \mathbb{E}_\pi\big[(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2}\big]\Big|^{4/3}\Big] \\
&\overset{(3)}{\le} \mathbb{E}_\pi\Big[\Big|(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2} - \mathbb{E}_\pi\big[(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2}\big]\Big|^2\Big]^{2/3} = \Big(\operatorname{var}_\pi\big((\frac{\mathrm{d}\mu}{\mathrm{d}\pi})^{3/2}\big)\Big)^{2/3}.
\end{aligned}$$

This proves the result. □

Theorem 2. *Suppose $\chi^2(\cdot \,\|\, \pi)$ satisfies the following Łojasiewicz inequality:*

$$\chi^2(\mu \,\|\, \pi)^{2/q} \le 4C_{\mathsf{PL}}\, \mathbb{E}_\mu\big[\|\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^2\big], \qquad \forall \mu \ll \pi, \tag{8}$$

for some $q \in (1, \infty)$. Then, π satisfies the chi-squared transportation inequality

$$W_2^2(\mu, \pi) \le p^2 C_{\mathsf{PL}}\, \chi^2(\mu \,\|\, \pi)^{2/p}, \qquad \forall \mu \ll \pi,$$

where $1/p + 1/q = 1$.

Proof. The proof follows [OV00]. Take a path $(\mu_t)_{t\geq 0}$ starting at some $\mu_0 = \mu$ and following the W_2 gradient flow of the chi-squared divergence $\chi^2(\cdot \,\|\, \pi)$, that is,

$$\partial_t \mu_t = 2 \operatorname{div}\big(\mu_t \nabla \frac{\mu_t}{\pi}\big).$$

The existence of this gradient flow and the regularity required for the following computations can be justified by [OT11; OT13] and [AGS08, Theorem 11.2.1]. Denote by T_t the optimal transport map sending μ_t to μ_0. Then, the time derivative of the squared Wasserstein distance can be computed as in [AGS08, Corollary 10.2.7] to be

$$\partial_t W_2^2(\mu_0, \mu_t) = -4\, \mathbb{E}_{\mu_t} \big\langle \nabla \frac{\mu_t}{\pi}, T_t - \mathrm{id} \big\rangle \leq 4 W_2(\mu_0, \mu_t)\, \mathbb{E}_{\mu_t} \big\| \nabla \frac{\mu_t}{\pi} \big\|,$$

where we apply the Cauchy-Schwarz and Jensen inequalities. It yields

$$\partial_t W_2(\mu_0, \mu_t) \leq 2\, \mathbb{E}_{\mu_t} \big\| \nabla \frac{\mu_t}{\pi} \big\|.$$

Also, the chi-squared divergence satisfies

$$\partial_t \chi^2(\mu_t \,\|\, \pi) = -4\, \mathbb{E}_{\mu_t} \big[\big\| \nabla \frac{\mu_t}{\pi} \big\|^2 \big].$$

Using the assumption (8),

$$\partial_t [\chi^2(\mu_t \,\|\, \pi)^{1/p}] = \frac{\partial_t \chi^2(\mu_t \,\|\, \pi)}{p \chi^2(\mu_t \,\|\, \pi)^{1/q}} = -\frac{4}{p \chi^2(\mu_t \,\|\, \pi)^{1/q}}\, \mathbb{E}_{\mu_t} \big[\big\| \nabla \frac{\mu_t}{\pi} \big\|^2 \big] \leq -\frac{2}{p\sqrt{C_{\mathsf{PL}}}}\, \mathbb{E}_{\mu_t} \big\| \nabla \frac{\mu_t}{\pi} \big\|.$$

If we define

$$g(t) := W_2(\mu_0, \mu_t) + p\sqrt{C_{\mathsf{PL}}}\, \chi^2(\mu_t \,\|\, \pi)^{1/p},$$

we have proved that

$$g' \leq 0.$$

Since $g(0) = p\sqrt{C_{\mathsf{PL}}}\, \chi^2(\mu_0 \,\|\, \pi)^{1/p}$ and $\lim_{t\to\infty} g(t) = W_2(\mu, \pi)$, we have shown a transport inequality

$$W_2^2(\mu, \pi) \leq p^2 C_{\mathsf{PL}}\, \chi^2(\mu \,\|\, \pi)^{2/p}. \qquad \square$$

Theorem 3. *Let π be a distribution on $\mathbb{R}^d$ with finite Poincaré constant $C_{\mathsf{P}} > 0$. Then for any measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, it holds*

$$W_2^2(\mu, \pi) \leq 9 C_{\mathsf{P}} \sqrt{\chi^2(\mu \,\|\, \pi)}\,.$$

Proof. The inequality follows immediately from the two preceding results. $\square$

Remark 1. Transportation-cost inequalities for Rényi divergences were also studied in [Din14; BD15].

C Additional choices for the mirror map

We extend our results to other choices of the mirror map ϕ that serve as proxies for V and that also lead to exponential convergence of MLD.

The first result below is useful in situations when there exists a strictly convex mirror map ϕ such $\nabla\phi$ is easier to invert than ∇V. It ensures exponential ergodicity of (MLD) when $\nabla^2 V$ dominates $\nabla^2 \phi$ in the sense of the Loewner order.

Corollary 5. *Suppose that π is strictly log-concave and that $\nabla^2\phi \preceq C \nabla^2 V$, where $\preceq$ denotes the Loewner order. Then, the law $(\mu_t)_{t\geq 0}$ of MLD satisfies*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \,\|\, \pi),\ \chi^2(\mu_t \,\|\, \mu),\ \frac{1}{2C_{\mathsf{P}}} W_2^2(\mu_t, \pi) \leq e^{-\frac{2t}{C}} \chi^2(\mu_0 \,\|\, \pi).$$

Proof. The assumption implies

$$C\, \mathbb{E}_\pi \langle \nabla f, (\nabla^2 \phi)^{-1} \nabla f \rangle \geq \mathbb{E}_\pi \langle \nabla f, (\nabla^2 V)^{-1} \nabla f \rangle \geq \mathrm{var}_\pi\, f,$$

where again we apply the Brascamp-Lieb inequality. This verifies (MP) with constant $C_{\mathsf{MP}} = C$. $\square$

Our second result does not require π to be log-concave but only that it is close to a strictly log-concave distribution $\widetilde{\pi}$ in the following sense: the density of π with respect to $\widetilde{\pi}$ is uniformly bounded away from 0 and ∞.

Corollary 6. *Suppose that $\widetilde{\pi} = \exp(-\widetilde{V})$ is strictly log-concave and suppose that π has density ρ w.r.t. $\widetilde{\pi}$. Let $M := (\sup \rho)/(\inf \rho)$. Then, the law $(\mu_t)_{t\geq 0}$ of* MLD *with mirror map $\phi = \widetilde{V}$ and target density π satisfies*

$$2\|\mu_t - \pi\|_{\mathrm{TV}}^2,\ H^2(\mu_t, \pi),\ D_{\mathrm{KL}}(\mu_t \parallel \pi),\ \chi^2(\mu_t \parallel \mu),\ \frac{1}{2C_{\mathsf{P}}M} W_2^2(\mu_t, \pi) \leq e^{-\frac{2t}{M}} \chi^2(\mu_0 \parallel \pi),$$

where C_{P} is the Poincaré constant of $\widetilde{\pi}$.

Proof. It is standard that the Poincaré inequality (P), and the mirror Poincaré inequality (MP), are stable under bounded perturbations of the measure. It implies that π satisfies a Poincaré inequality with constant $C_{\mathsf{P}}M$, and a mirror Poincaré inequality with constant M. We prove the latter statement for completeness; for the former statement, see [Han16, Problem 3.20].

Observe that

$$\int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1} \nabla f \rangle \, \mathrm{d}\pi = \int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1} \nabla f \rangle \, \frac{\mathrm{d}\pi}{\mathrm{d}\widetilde{\pi}} \, \mathrm{d}\widetilde{\pi} \geq (\inf \rho) \int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1} \nabla f \rangle \, \mathrm{d}\widetilde{\pi}$$

and

$$\begin{aligned} \mathrm{var}_{\widetilde{\pi}} f &= \inf_{m \in \mathbb{R}^d} \int \|f - m\|^2 \, \mathrm{d}\widetilde{\pi} = \inf_{m \in \mathbb{R}^d} \int \|f - m\|^2 \, \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi} \, \mathrm{d}\pi \\ &\geq \frac{1}{\sup \rho} \inf_{m \in \mathbb{R}^d} \int \|f - m\|^2 \, \mathrm{d}\pi = \frac{1}{\sup \rho} \mathrm{var}_{\pi} f. \end{aligned}$$

Combining these inqualities with the Brascamp-Lieb inequality for $\widetilde{\pi}$,

$$\int \langle \nabla f, (\nabla^2 \widetilde{V})^{-1} \nabla f \rangle \, \mathrm{d}\widetilde{\pi} \geq \mathrm{var}_{\widetilde{\pi}} f,$$

yields (MP) with constant $C_{\mathsf{MP}} = M$. □

D Stability in KL with respect to exp-concave perturbations

The following lemma quantifies the approximation error of replacing π by π_β in Section 4.2 and, more generally provides a simple bound to control the KL divergence between a log-concave distribution and its perturbation by a ν-exp-concave barrier function. Its proof uses crucially displacement convexity of the KL divergence to a log-concave measure [Vil03, §5], and it can be viewed as the sampling analogue of [Nes04, (4.2.17)].

Recall that b is *ν-exp-concave* if the mapping $\exp(-\nu^{-1} b)$ is concave.

Lemma 1. *Let π be a log-concave distribution on a convex set $\mathcal{K} \subset \mathbb{R}^d$. Fix $\nu > 0$, and let $\widetilde{\pi}$ have density $\exp(-b)$ with respect to π, where $b : \mathcal{K} \to \mathbb{R}$ is ν-exp-concave. Then it holds that*

$$D_{\mathrm{KL}}(\widetilde{\pi} \parallel \pi) \leq \nu \,.$$

Proof. On $\operatorname{int} \mathcal{K}$, we have

$$-\nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi} = \nabla b. \tag{9}$$

The measure π is log-concave, so by displacement convexity of entropy [AGS08, Theorem 9.4.11] and the "above-tangent" formulation of convexity [Vil03, Proposition 5.29], we have

$$0 = D_{\mathrm{KL}}(\pi \parallel \pi) \geq D_{\mathrm{KL}}(\widetilde{\pi} \parallel \pi) + \mathbb{E}\big\langle \nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi}(\widetilde{X}), X - \widetilde{X} \big\rangle,$$

where $(X, \widetilde{X})$ are optimally coupled for π and $\widetilde{\pi}$. If we rearrange this inequality and use the identities in (9), we get

$$D_{\mathrm{KL}}(\widetilde{\pi} \parallel \pi) \leq -\,\mathbb{E}\big\langle \nabla \ln \frac{\mathrm{d}\widetilde{\pi}}{\mathrm{d}\pi}(\widetilde{X}), X - \widetilde{X} \big\rangle = \mathbb{E}\langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle \,. \tag{10}$$

We now use the fact that b is ν-exp-concave. To that end, define the convex function

$$\varphi(t) = -\exp\big(-\frac{1}{\nu} b(\widetilde{X} + t\,(X - \widetilde{X}))\big), \qquad t \in [0,1]\,.$$

By convexity, we have

$$\varphi'(0) \cdot (1-0) \le \varphi(1) - \varphi(0) \le -\varphi(0) = \exp\big(-\frac{1}{\nu} b(\widetilde{X})\big).$$

Since

$$\varphi'(0) = \frac{1}{\nu} \exp\big(-\frac{1}{\nu} b(\widetilde{X})\big)\, \langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle\,,$$

the above inequality reads $\langle \nabla b(\widetilde{X}), X - \widetilde{X} \rangle \le \nu$, which completes the proof together with (10). □

Remark 2. It is known that given any convex body $\mathcal{C} \subset \mathbb{R}^d$, there exists a standard self-concordant ν^{-1}-exp-concave barrier with $\nu \le d$ [NN94; BE15; TY18].

E Numerical experiments

In this section, we gather additional details and figures to support our numerical experiments. First, in Section E.1, we display the samples from our Gaussian experiment. Then, Section E.2 gives details of the Bayesian logistic regression experiment displayed in Figure 1 and shows the effect of varying step size. Section E.3 gives details of sampling from an ill-conditioned convex set. Finally, Section E.4 shows an experiment where we use the NLA and a Mirror-Langevin Algorithm MLA to approximately sample from a degenerate log-concave distribution.

E.1 Sampling from a Gaussian distribution

We display some supplementary experiments for the elliptically symmetric scatter family example of Section 5. First, we repeat the example in Figure 4 for the simpler case of the Gaussian distribution ($\gamma = 1$) on $\mathbb{R}^{100}$ with the same scatter matrix $\Sigma = \mathrm{diag}(1, 2, \dots, 100)$ in Figure 5. We again see the superiority of NLA over the Unadjusted Langevin Algorithm (ULA) [DM+19] and the Tamed Unadjusted Langevin Algorithm (TULA) [Bro+19]. Here and in Section 5 the additional parameter of TULA (denoted γ in [Bro+19]) is chosen equal to .1.

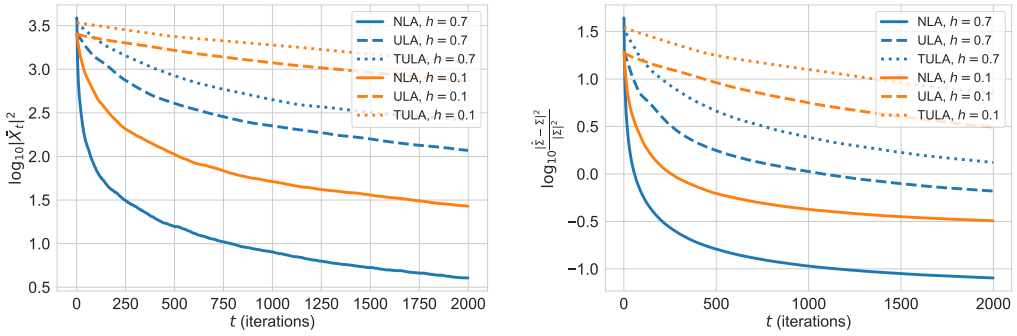

Figure 5: We display convergence of the various algorithms for an ill-conditioned Gaussian distribution, with $d = 100$ and $\Sigma = \mathrm{diag}(1, 2, \dots, 100)$. Left: error is the squared distance from 0. Right: error is the relative distance between scatter matrices. As in the experiment displayed in Figure 4, NLA rapidly converges both in terms of location and scale for large step sizes.

We also display some samples from the Gaussian experiment of Figure 5 in Figure 6. NLA maintains good performance for a wide range of step-size choices, while ULA and TULA require a small step size to accurately sample from the target distribution. In fact, even with a small step size, ULA and TULA often jump to small probability regions, while NLA avoids these regions even for large step sizes.

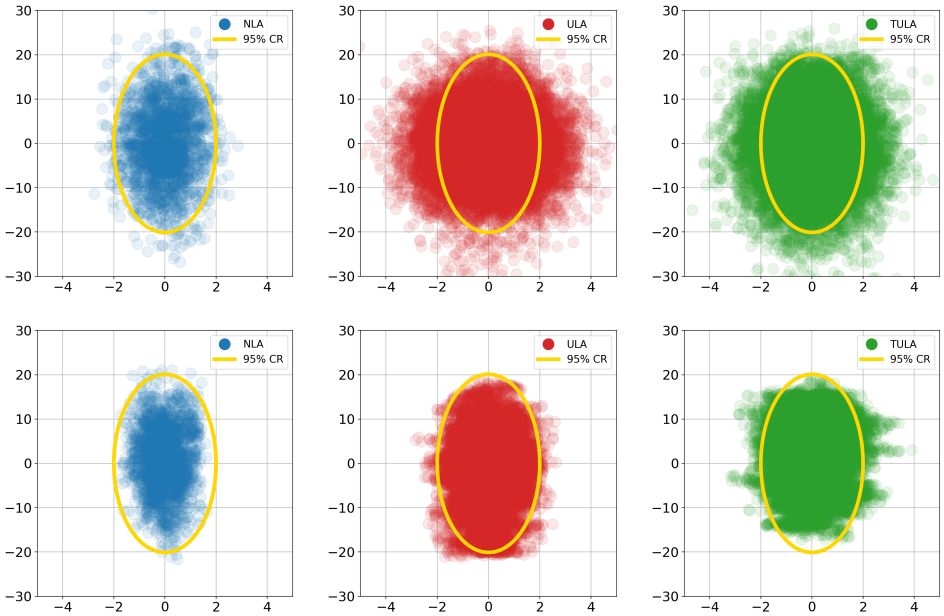

Figure 6: Samples from NLA, ULA, and TULA for the ill-conditioned Gaussian example of Figure 5, with $\Sigma = \mathrm{diag}(1, 2, \ldots, 100)$. We display the projection onto the first (least spread) and last (most spread) population principal components, along with the projection of a 95% confidence region. Top: the step size for all algorithms is $h = 0.7$, Bottom: the step size for all algorithms is $h = 0.05$.

E.2 Bayesian logistic regression

We give details for the two-dimensional Bayesian logistic regression example in Figure 1. In the Bayesian logistic regression model, covariates are drawn as $X_i \sim \mathcal{N}(0, \mathrm{diag}(10, 0.1))$, the response variables are $Y_i \sim \mathsf{Ber}(\mathrm{logit}(\theta^\top X_i))$, and the parameters θ have a $\mathcal{N}(0, 10 I_2)$ prior. We consider using NLA to sample from the posterior distribution of θ given the observations (X_i, Y_i), $i = 1, \ldots, n$, which is

$$\pi(\theta) \propto \exp\Big[-\frac{1}{20}\|\theta\|^2 + \sum_{i=1}^{n} Y_i \theta^\top X_i - \ln(1 + e^{\theta^\top X_i})\Big],$$

which is strongly log-concave. While the gradient of the potential is invertible, it has no closed-form, and so in our experiments we invert it numerically by solving $\nabla V^\star(y) = \mathrm{argmax}_{x \in \mathbb{R}^d} \{\langle x, y\rangle - V(x)\}$ with Newton's method. We find that, with a warm start from the current iterate X_t, it suffices to run Newton's method for a small number of iterations to approximately invert the gradient.

For the purposes of this experiment, we generate 100 samples $X_i \sim \mathcal{N}(0, \mathrm{diag}(10, 0.1))$ and $Y_i \sim \mathsf{Ber}(\mathrm{logit}(\theta^{\star\top} X_i))$, where we set $\theta^\star = (1, 1)$.

We display the result for various sampling algorithms in Figure 1. All algorithms are implemented with $h = 0.1$ and a burn-in time of 10^4 steps. This example shows the advantage of taking a large step-size with NLA in this ill-conditioned model, while ULA and TULA create samples that are overdispersed. In Figure 7, we also show the effect of decreasing step size in this example. In this case, we see that ULA and TULA still step into low probability regions or fail to explore the underlying density well. On the other hand, NLA remains constrained in the high probability region.

E.3 Uniform sampling on a convex body

This section contains details for the simulations in Figure 3. We sample from the uniform distribution on the rectangle $[-0.01, 0.01] \times [-1, 1]$ using NLA, PLA, and the Metropolis-Adjusted Langevin Algorithm (MALA) [Dwi+19]. PLA and MALA target the uniform distribution directly. NLA samples from an approximate distribution, given in Section 4.2. The step sizes are chosen as $h = 10^{-5}$ for NLA and PLA and $h = 0.01$ for MALA. The step sizes for PLA and MALA are tuned to allow the

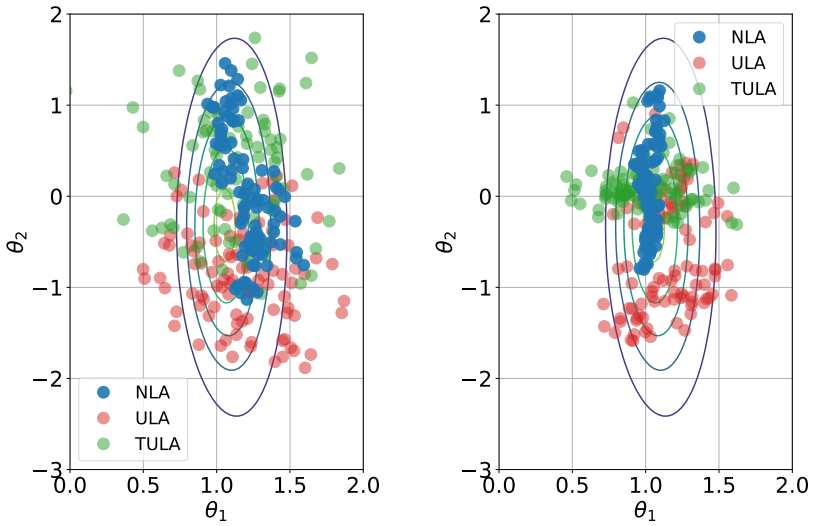

Figure 7: Samples from the posterior distribution of a Bayesian logistic regression model using one run of NLA, ULA, and TULA after a burn-in of 10^4. Left: large step size (all algorithms use $h = 0.05$); NLA remains within the high-density contours, while the ULA and TULA take steps into low-density areas. Right: small step size (all algorithms use $h = 0.01$); NLA explores the underlying distribution faster than its competitors.

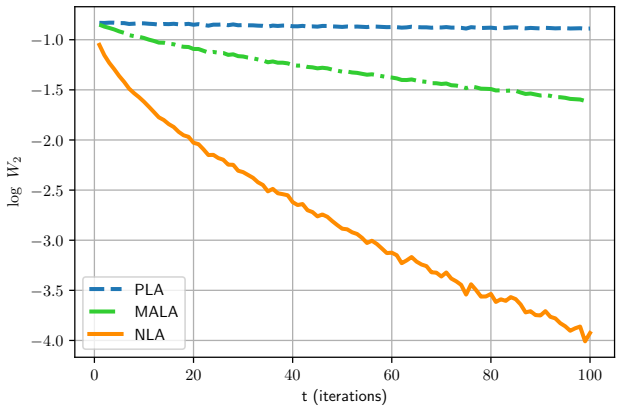

Figure 8: W_2 distance (on logarithmic scale) between the uniform distribution on the rectangle $[-0.01, 0.01] \times [-1, 1]$, and samples produced by NLA, PLA, and MALA.

algorithm to reach approximate stationarity in the fewest number of iterations. MALA can use a larger step size because it is unbiased (its stationary distribution coincides with the target distribution, due to the Metropolis-Hastings adjustment). On the other hand, samples from PLA tend to cluster around the boundary for larger step sizes, so we use a smaller step size for both PLA (and NLA for fair comparison).

To evaluate the performance of the algorithms, we estimate the 2-Wasserstein distance between the samples drawn by the algorithms and samples drawn from the uniform distribution on the rectangle; see Figure 8. We use the Sinkhorn distance ($\varepsilon = 0.01$) as an approximation for the 2-Wasserstein distance [Cut13; AWR17]. Specifically, we sample 1000 points in parallel, using the three algorithms of interest. At each iteration, we also draw 1000 points from the uniform distribution on the rectangle, and we compute the Sinkhorn distance between these points and the samples produced by the algorithms. The convergence estimates are averaged over 30 runs.

E.4 Approximate sampling from degenerate log-concave distributions

In this section, we explore further the problem of approximately sampling according to the measure $\pi(x) \propto \exp(-\|x\|)$ in $\mathbb{R}^2$ considered in Figure 2. To that end, we use the penalization strategy outlined in Section 4.1 and sample instead from the strongly log-concave measure $\pi_\beta(x) \propto \exp(-\|x\| - \beta\|x - \mathbf{1}\|^2)$ as in Corollary 2, where $\beta = 0.0005$, using discretizations of either NLD or MLD with a customized mirror map. Here, $\mathbf{1}$ is the vector of all ones, which simulates the effect of not knowing the true mean.

We initialize all algorithms with a random point X_0 with $\|X_0\| = 1000$. The initialization is intentionally chosen so that the gradients of the potential at initialization are extremely small. In these circumstances, we expect ULA to mix slowly.

Through this experiment, we demonstrate two empirical observations:

1. Initially, the iterates of NLA converge extremely rapidly to the vicinity of the origin. This suggests that NLA can be useful for initializing other sampling algorithms in highly ill-conditioned settings.
2. However, once the iterates of NLA are near the origin, NLA becomes unstable. Specifically, since the Hessian of the potential degenerates rapidly near 0, the iterates of NLA occasionally make large jumps away from 0. This is due to the fact that the Hessian of $V(x) = \|x\| + \beta\|x - \mathbf{1}\|^2$ is given by
$$\nabla^2 V(x) = \frac{1}{\|x\|}\Big[I_2 - \Big(\frac{x}{\|x\|}\Big)\Big(\frac{x}{\|x\|}\Big)^\top\Big] + 2\beta I_2 \tag{11}$$
which blows up to infinity around $x = 0$. We remark that Newton's method in optimization can also exhibit unstable behavior [CGT00; NP06], so this phenomenon is not unexpected. To rectify this behavior, we also consider the Euler discretization of MLD, which we call MLA (see below). We demonstrate that with an appropriate choice of mirror map, the iterates of MLA are stable, yet still enjoy faster convergence than ULA.

Now we proceed to the details of the experiment. We compare four different methods for sampling from this distribution: NLA, ULA, TULA, and the mirror-Langevin Algorithm (MLA)
$$\nabla\phi(X_{k+1}) = \nabla\phi(X_k) - h\nabla V(X_k) + \sqrt{2h}\,[\nabla^2\phi(X_k)]^{1/2}\xi_k, \tag{MLA}$$
with mirror map $\phi(x) = \|x\|^{3/2}$ and potential $V(x) = \|x\| + \beta\|x - \mathbf{1}\|^2$. Notice that this mirror map corresponds to that used in the generalized Gaussian case of Section 5.

In Figure 9, we display the results of the first 1000 iterations of the four algorithms. In this stage of the experiment, we observe rapid convergence of NLA towards the origin (around which the mass is concentrated), and MLA also exhibits faster convergence than ULA and TULA. However, already in Figure 9 (Right) we observe the instability of NLA witnessed through large jumps of the iterates.

Next, in Figure 10, we treat the samples from the first 1000 iterations as burn-in, and we look at the performance of the next 1000 samples. Here we see that the flexible framework of the more general MLD allows us to design algorithms which can outperform NLA with superior stability in specific scenarios.

Recall that the Hessian of the potential V is given in (11) while the potential of the mirror map ϕ is given by
$$\nabla^2\phi(x) = \frac{3}{2\|x\|^{1/2}}\Big[I_2 - \frac{3}{4}\Big(\frac{x}{\|x\|}\Big)\Big(\frac{x}{\|x\|}\Big)^\top\Big].$$
From these expressions, it can be checked that Corollary 5 holds with $C \le 3/(4\sqrt{2\beta})$. On the other hand, the measure π_β satisfies a Poincaré inequality (P) with constant $C_{\mathsf{P}} \le 1/(2\beta)$. Heuristically, we therefore expect the mixing time of ULA to scale as $O(\beta^{-1})$, and the mixing time of MLA to scale as $O(\beta^{-1/2})$, which provides an explanation for the rates of convergence observed in Figure 9. In comparison, the mixing time of NLA is scale-invariant, i.e. $O(1)$, as we demonstrated in Corollary 1, as witnessed by the initial rapid convergence in Figure 9.

As mentioned in our open questions, this points to the intriguing possibility of developing more stable variants of NLA, which would mirror the development of such strategies for Newton's method [CGT00; NP06].

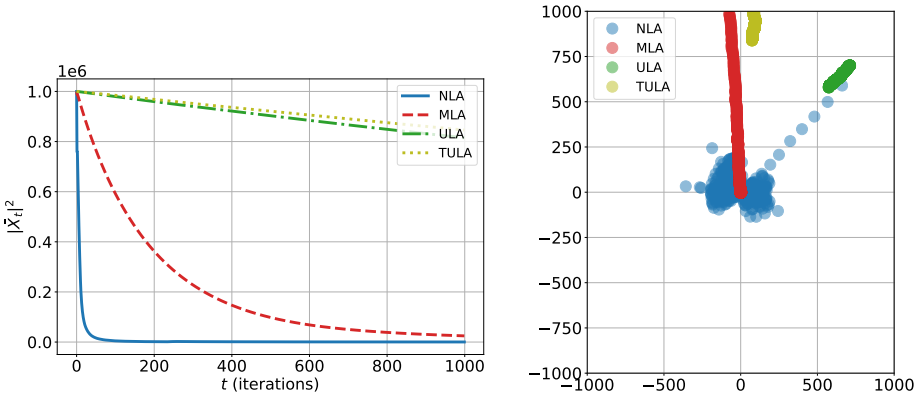

Figure 9: First stage of the experiment. Left: We plot the norm of the running mean versus the iteration number for the target measure $\pi_\beta(x) \propto \exp(-\|x\| - 0.0005\|x - \mathbf{1}\|^2)$. Right: We display the corresponding samples.

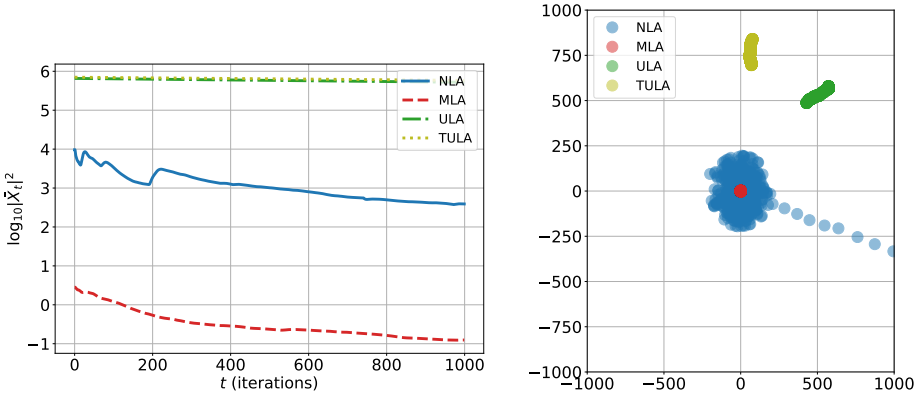

Figure 10: Second stage of the experiment. In this stage, we treat the 1000 samples from the first stage of the experiment as burn-in and look at the performance of the next 1000 samples. Left: We plot the logarithm of the norm of the running mean versus iteration. Right: We again display the corresponding samples.

F Broader impact

The sampling algorithms designed in this paper have the potential to improve a wide variety of Bayesian methods and therefore have an indirect impact on various domains such as health and medicine where such methods are pervasive. Sampling algorithms are also used for the generation of automated spam messages, which have potentially negative effects on society. Since this paper is primarily focused on theory, these questions are not addressed here.

G Acknowledgments

Philippe Rigollet was supported by NSF awards IIS-1838071, DMS-1712596, DMS-TRIPODS-1740751. Sinho Chewi and Austin Stromme were supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. Thibaut Le Gouic was supported by ONR grant N00014-17-1-2147 and NSF IIS-1838071.

We thank the reviewers for very helpful suggestions regarding the presentation of the paper.

References

[AB15] D. Alonso-Gutiérrez and J. Bastero. *Approaching the Kannan-Lovász-Simonovits and variance conjectures*. Vol. 2131. Lecture Notes in Mathematics. Springer, Cham, 2015, pp. x+148.

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[AWR17] J. Altschuler, J. Weed, and P. Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 2017, pp. 1961–1971.

[BC12] S. Bubeck and N. Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122.

[BCG08] D. Bakry, P. Cattiaux, and A. Guillin. "Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré". In: *Journal of Functional Analysis* 254.3 (2008), pp. 727–759.

[BD15] S. G. Bobkov and Y. Ding. "Optimal transport and Rényi informational divergence". In: *Electron. Commun. Probab.* 20 (2015), no. 4, 12.

[BE15] S. Bubeck and R. Eldan. "The entropic barrier: a simple and optimal universal self-concordant barrier". In: *Conference on Learning Theory*. 2015, pp. 279–279.

[BEL18] S. Bubeck, R. Eldan, and J. Lehec. "Sampling from a log-concave distribution with projected Langevin Monte Carlo". In: *Discrete & Computational Geometry* 59.4 (2018), pp. 757–783.

[Ber18] E. Bernton. "Langevin Monte Carlo and JKO splitting". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1777–1798.

[BGG12] F. Bolley, I. Gentil, and A. Guillin. "Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations". In: *Journal of Functional Analysis* 263.8 (2012), pp. 2430–2457.

[BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.

[BL00] S. G. Bobkov and M. Ledoux. "From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities". In: *Geom. Funct. Anal.* 10.5 (2000), pp. 1028–1052.

[BL06] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization*. Second. Vol. 3. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Theory and examples. Springer, New York, 2006, pp. xii+310.

[BL76] H. J. Brascamp and E. H. Lieb. "On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation". In: *J. Functional Analysis* 22.4 (1976), pp. 366–389.

[Bob99] S. G. Bobkov. "Isoperimetric and analytic inequalities for log-concave probability measures". In: *The Annals of Probability* 27.4 (1999), pp. 1903–1921.

[Bro+19] N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. "The tamed unadjusted Langevin algorithm". In: *Stochastic Processes and their Applications* 129.10 (2019), pp. 3638–3663.

[Bub15] S. Bubeck. "Convex optimization: algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[BV05] F. Bolley and C. Villani. "Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities". In: *Annales de La Faculté Des Sciences de Toulouse: Mathématiques*. Vol. 14. 3. 2005, pp. 331–352.

[CB18] X. Cheng and P. Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Algorithmic Learning Theory 2018*. Vol. 83. Proc. Mach. Learn. Res. (PMLR). Proceedings of Machine Learning Research PMLR, 2018, p. 26.

[CG09] P. Cattiaux and A. Guillin. "Trends to equilibrium in total variation distance". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 45.1 (2009), pp. 117–145.

[CGT00] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. Vol. 1. SIAM, 2000.

[Che+18] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. "Underdamped Langevin MCMC: A non-asymptotic analysis". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 300–323.

[Che+19] X. Cheng, D. Yin, P. L. Bartlett, and M. I. Jordan. "Quantitative W_1 convergence of Langevin-like stochastic processes with non-convex potential and state-dependent noise". In: *arXiv e-prints*, arXiv:1907.03215 (July 2019).

[Che+20] S. Chewi, T. Maunu, P. Rigollet, and A. Stromme. "Gradient descent algorithms for Bures-Wasserstein barycenters". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 1276–1304.

[CL89] M.-F. Chen and S.-F. Li. "Coupling methods for multidimensional diffusion processes". In: *The Annals of Probability* (1989), pp. 151–177.

[CLL19] Y. Cao, J. Lu, and Y. Lu. "Exponential decay of Rényi divergence under Fokker-Planck equations". In: *J. Stat. Phys.* 176.5 (2019), pp. 1172–1184.

[Cor17] D. Cordero-Erausquin. "Transport inequalities for log-concave measures, quantitative forms, and applications". In: *Canad. J. Math.* 69.3 (2017), pp. 481–501.

[Cut13] M. Cuturi. "Sinkhorn distances: lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., 2013, pp. 2292–2300.

[Dal17a] A. Dalalyan. "Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, 2017, pp. 678–689.

[Dal17b] A. S. Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

[Din14] Y. Ding. "Wasserstein-divergence transportation inequalities and polynomial concentration inequalities". In: *Statist. Probab. Lett.* 94 (2014), pp. 77–85.

[Din15] Y. Ding. "A note on quadratic transportation and divergence inequality". In: *Statist. Probab. Lett.* 100 (2015), pp. 115–123.

[DK19] A. S. Dalalyan and A. Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *Stoch. Proc. Appl.* 129.12 (2019), pp. 5278–5311.

[DM+19] A. Durmus, É. Moulines, et al. "High-dimensional Bayesian inference via the unadjusted Langevin algorithm". In: *Bernoulli* 25.4A (2019), pp. 2854–2882.

[DM15] A. Durmus and É. Moulines. "Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm". In: *Statistics and Computing* 25.1 (Jan. 2015), pp. 5–19.

[DM17] A. Durmus and É. Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *Ann. Appl. Probab.* 27.3 (2017), pp. 1551–1587.

[DMM19] A. Durmus, S. Majewski, and B. Miasojedow. "Analysis of Langevin Monte Carlo via convex optimization". In: *J. Mach. Learn. Res.* 20 (2019), Paper No. 73, 46.

[DR20] A. S. Dalalyan and L. Riou-Durand. "On sampling from a log-concave density using kinetic Langevin diffusions". In: *Bernoulli* 26.3 (2020), pp. 1956–1988.

[DRK19] A. S. Dalalyan, L. Riou-Durand, and A. Karagulyan. "Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets". In: *arXiv e-prints*, arxiv:1906.08530 (June 2019).

[Dwi+19] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. "Log-concave sampling: Metropolis-Hastings algorithms are fast". In: *Journal of Machine Learning Research* 20.183 (2019), pp. 1–42.

[Ebe16] A. Eberle. "Reflection couplings and contraction rates for diffusions". In: *Probability Theory and Related Fields* 166.3-4 (2016), pp. 851–886.

[FKP94] A. Frieze, R. Kannan, and N. Polson. "Sampling from log-concave distributions". In: *The Annals of Applied Probability* 4.3 (1994), pp. 812–837.

[Gen08] I. Gentil. "From the Prékopa-Leindler inequality to modified logarithmic Sobolev inequality". In: *Ann. Fac. Sci. Toulouse Math. (6)* 17.2 (2008), pp. 291–308.

[Goo+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. 2014, pp. 2672–2680.

[Han16] R. van Handel. *Probability in high dimension*. Lecture Notes (Princeton University). 2016.

[Hsi+18] Y.-P. Hsieh, A. Kavis, P. Rolland, and V. Cevher. "Mirrored Langevin dynamics". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2878–2887.

[JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. "The variational formulation of the Fokker-Planck equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.

[KLS95] R. Kannan, L. Lovász, and M. Simonovits. "Isoperimetric problems for convex bodies and a localization lemma". In: *Discrete Comput. Geom.* 13.3-4 (1995), pp. 541–559.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2016, pp. 795–811.

[KS91] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470.

[Led18] M. Ledoux. *Remarks on some transportation cost inequalities*. 2018.

[Liu20] Y. Liu. "The Poincaré inequality and quadratic transportation-variance inequalities". In: *Electron. J. Probab.* 25 (2020), Paper No. 1, 16.

[LL08] C. Le Bris and P.-L. Lions. "Existence and uniqueness of solutions to Fokker-Planck type equations with irregular coefficients". In: *Comm. Partial Differential Equations* 33.7-9 (2008), pp. 1272–1317.

[Loj63] S. Lojasiewicz. "Une propriété topologique des sous-ensembles analytiques réels". In: *Les équations aux dérivées partielles* 117 (1963), pp. 87–89.

[LTV20] A. Laddha, Y. Tat Lee, and S. Vempala. "Strong self-concordance and sampling". In: *STOC* (2020).

[LV07] L. Lovász and S. Vempala. "The geometry of logconcave functions and sampling algorithms". In: *Random Structures & Algorithms* 30.3 (2007), pp. 307–358.

[LV17] Y. T. Lee and S. S. Vempala. "Eldan's stochastic localization and the KLS hyperplane conjecture: an improved lower bound for expansion". In: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 998–1007.

[LV18] Y. T. Lee and S. S. Vempala. "Stochastic localization + Stieltjes barrier = tight bound for log-Sobolev". In: *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2018, pp. 1122–1129.

[Ma+19] Y.-A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett, and M. I. Jordan. "Is there an analog of Nesterov acceleration for MCMC?" In: *arXiv e-prints*, arXiv:1902.00996 (Feb. 2019).

[Mar+12] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion". In: *SIAM J. Sci. Comput.* 34.3 (2012), A1460–A1487.

[Mou+19] W. Mou, N. Flammarion, M. J. Wainwright, and P. L. Bartlett. "Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity". In: *arXiv e-prints*, arXiv:1907.11331 (July 2019).

[MT09] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. 2nd. USA: Cambridge University Press, 2009.

[Nea12] R. Neal. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo* (June 2012).

[Nes04] Y. Nesterov. *Introductory lectures on convex optimization*. Vol. 87. Applied Optimization. A basic course. Kluwer Academic Publishers, Boston, MA, 2004, pp. xviii+236.

[NJ79] A. S. Nemirovskii and D. B. Judin. *Complexity of problems and efficiency of optimization methods*. 1979.

[NN94] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Vol. 13. SIAM Studies in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994, pp. x+405.

[NP06] Y. Nesterov and B. T. Polyak. "Cubic regularization of Newton method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205.

[OT11] S.-i. Ohta and A. Takatsu. "Displacement convexity of generalized relative entropies". In: *Adv. Math.* 228.3 (2011), pp. 1742–1787.

[OT13] S.-i. Ohta and A. Takatsu. "Displacement convexity of generalized relative entropies. II". In: *Comm. Anal. Geom.* 21.4 (2013), pp. 687–785.

[OV00] F. Otto and C. Villani. "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality". In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400.

[RC04] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

[Roc97] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Reprint of the 1970 original, Princeton Paperbacks. Princeton University Press, Princeton, NJ, 1997, pp. xviii+451.

[RRT17] M. Raginsky, A. Rakhlin, and M. Telgarsky. "Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, July 2017, pp. 1674–1703.

[San17] F. Santambrogio. "{Euclidean, metric, and Wasserstein} gradient flows: an overview". In: *Bulletin of Mathematical Sciences* 7.1 (2017), pp. 87–154.

[Sim+16] U. Simsekli, R. Badeau, A. T. Cemgil, and G. Richard. "Stochastic quasi-Newton Langevin Monte Carlo". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 642–651.

[SKL20] A. Salim, A. Korba, and G. Luise. "Wasserstein proximal gradient". In: *arXiv e-prints*, arxiv:2002.03035 (Feb. 2020).

[Tsy09] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214.

[TY18] Y. Tat Lee and M.-C. Yue. "Universal barrier is n-self-concordant". In: *arXiv e-prints*, arxiv:1809.03011 (Sept. 2018).

[Vil03] C. Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.

[Vil09] C. Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973.

[VW19] S. Vempala and A. Wibisono. "Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8094–8106.

[Wib18] A. Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem". In: *Conference on Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2093–3027.

[Wib19] A. Wibisono. "Proximal Langevin algorithm: rapid convergence under isoperimetry". In: *arXiv e-prints*, arxiv:1911.01469 (Nov. 2019).

[WL20] Y. Wang and W. Li. "Information Newton's flow: second-order optimization method in probability space". In: *arXiv e-prints*, arxiv:2001.04341 (Jan. 2020).

[Zha+20] K. S. Zhang, G. Peyré, J. Fadili, and M. Pereyra. "Wasserstein control of mirror Langevin Monte Carlo". In: *arXiv e-prints*, arxiv:2002.04363 (Feb. 2020).

[ZWG13] T. Zhang, A. Wiesel, and M. S. Greco. "Multivariate generalized Gaussian distribution: convexity and graphical models". In: *IEEE Transactions on Signal Processing* 61.16 (2013), pp. 4141–4148.